**Bioinformatics Exercise 2:** Query a public database

*Examination of a GenBank entry*

* Go to the [Genbank website](http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide)
* Search for “DMD AND Homo sapiens”
* Select the RefSeq tab (left side)
* Select transcript variant Dp427m, (accession code NM\_004006)
* Explore the information, e.g. gene, coding sequence (CDS) , PubMed
* Find the variation in region 3713. (scroll way down) **What effect has this variant?**

*Retrieve a sequence from the database and store it locally*

The "fasta" format is a format for sequences, which can be read by many programs.

* Search in the Genbank database for 'human mRNA HBA2' (NM\_000517) Tip: you can search on accession code
* Set Display (top left of the page) to Fasta
* Save as a text file, save as NM\_000517.txt (This is only to show how you can save information from GenBank, we do not need this sequence anymore)

**What are the characteristics of the FASTA format?**

**Exercise 3:** Compare two sequences to identify mutations

* Read more about the Sickle cell anemia gene [here](http://ghr.nlm.nih.gov/condition/sickle-cell-disease) and [here.](http://ghr.nlm.nih.gov/gene/HBB)
* **Which mutation causes Sickle cell disease?**

The bl2seq program (BLAST) is used to align two sequences against each other instead of against the entire database. The sequence can be pasted into the boxes or the accession codes (unique identifier for a sequence) can be entered as input.

* Go to the [BLAST website](http://www.ncbi.nlm.nih.gov/BLAST/). Determine the position of the HBS mutation with the program "Align two sequences (bl2seq)" (In the "Specialized Blast" box)
* Type "NM\_000518" in the box for sequence1, this is the mRNA sequence of normal HBB.
* Type "M25113" (mRNA of HBS) in the second box.
* Click on "Blast" and examine the output.
* At the top of the page you can alter the 'formatting options'. Tick the "CDS feature" box and click on "Reformat" to see the protein translations of the sequences.

The sequences are retrieved from the nucleotide database (GenBank) and are placed above each other. This is called an alignment. Matching nucleotides are indicated by a pipe-sign: |. Mismatching nucleotides will not have a pipe-sign. Insertions/deletions are annotated as a dash (-) in one of the sequences. Below the alignment the protein sequence is placed when the annotation of one of the sequences is known. In this case the identifiers were given as input, so the bl2seq program could retrieve this information from GenBank.

* **Identify the mutation that causes the change in amino acid.**